Award Number: W81XWH-12-1-0106

TITLE: Use of a Novel Embryonic Mammary Stem Cell Gene Signature to Improve Human Breast Cancer Diagnostics and Therapeutic Decision Making

PRINCIPAL INVESTIGATOR: Geoffrey M. Wahl, Ph.D.

CONTRACTING ORGANIZATION: Salk Institute
La Jolla, CA 92037

REPORT DATE: October 2013

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE October 2013	**2. REPORT TYPE** Annual	**3. DATES COVERED** 30 Sep 2012 – 29 Sep 2013

4. TITLE AND SUBTITLE Use of a Novel Embryonic Mammary Stem Cell Gene Signature to Improve Human Breast Cancer Diagnostics and Therapeutic Decision Making

5a. CONTRACT NUMBER

5b. GRANT NUMBER W81XWH-12-1-0106

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Wahl, Geoffrey M.; Perou, Charles; Spike, Benjamin; Lasken, Roger

E-Mail: cperou@med.unc.edu

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Salk Institute
La Jolla, CA 92037

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT
Our major goals are to determine whether Fetal Mammary Stem Cell (fMaSC) signatures correlate with response to chemotherapy and metastasis in different breast cancer intrinsic subtypes (AIM1), and to develop single cell sequencing to produce highly refined fMaSC signatures (AIM2). Accomplishing these aims will enable us to: 1) better categorize distinct cell types within the fMaSC population, 2) identify biomarkers for prospective stem cell purification and in situ localization, and 3) identify candidate stem cell regulatory pathways that should reveal therapeutic targets and improved prognosticators and response biomarkers. In the most recent funding period, our bioinformatic analysis identified subsets of fMaSC signature genes that are coordinately expressed in archived human breast cancer gene expression data sets and assessed their prognostic and/or predictive power. We have thus far identified one subset exhibiting significant prognostic value distinct from existing and commonly used clinical variables in the preliminary data sets we have analyzed. We have also adapted a new microfluidics-based, single-cell capture and library preparation system to improve reproducibility in the generation of gene expression profiles from individual fMaSC. These advances provide proof of the principles underlying this grant and leave us well positioned to achieve its aims.

15. SUBJECT TERMS
Breast Cancer Prognosis, Mammary Stem Cells, Embryonic Development, Single Cell Transcriptomics

16. SECURITY CLASSIFICATION OF: Unclassified			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** USAMRMC
a. REPORT U	**b. ABSTRACT** U	**c. THIS PAGE** U	UU	14	**19b. TELEPHONE NUMBER** *(include area code)*

Table of Contents

Introduction:

This Idea Award Expansion proposed two aims to capitalize on the significant progress we made in the originally funded Idea Award. Studies performed with funding from the original Idea Award resulted in the first identification, isolation, and characterization of stem cells derived from the embryonic mouse mammary gland that are committed to a mammary fate. We showed that genes expressed in these Fetal Mammary Stem Cells (fMaSCs) are highly enriched in basal-like breast cancers.

Aim 1 of this Expansion Award proposed to derive more refined signatures from the fMaSC signature to see if they generate stronger predictors of chemotherapy response or metasis than afforded by currently available predictors derived from adult cell populations Our hypothesis was that the unique, biologically defined fMaSC cell population, might enable us to obtain useful prognostic signatures since they exhibit gene ontology pathways similar to those of aggressive tumor cells and their overall signature is enriched in late stage, basal-like breast cancers.

Aim 2 took on the significant challenge of using single cell RNA-sequencing to deconvolute the fMaSC population into its component cell types. Based on *in vitro* sphere formation and *in vivo* limiting dilution transplantation functional assays, we estimated the fMaSC population to be 10-20% pure. Therefore, we inferred that its gene expression signature represented the constellation of cell types present in our most purified fMaSC population. We therefore proposed to use single cell RNA-sequencing methods to evaluate the gene expression characteristics of individual cells we inferred to be stem cell candidates as they express both myoeptithelial (K14) and luminal (K8) cytokeratins, versus cells on their way to luminal or myoepithelial differentiation that express only K8 and luminal differentiation genes, or only K14 and myoepithelial differentiation genes. This was a very challenging objective as the technology for single cell RNA-seq and the bioinformatic tools for analyzing such data were just becoming available at the time we submitted this Idea Expansion proposal. Indeed, as summarized below, we encountered significant challenges that slowed our progress, but have now overcome many of them and are well positioned to complete this Aim in the coming year.

Body:

Aim 1

A major goal of this Aim was to determine whether the Fetal Mammary Stem Cell (fMaSC) signature correlates with response to chemotherapy in different intrinsic subtypes of breast cancer. We first obtained genomic data from five independent studies that used fluorescence activated cell sorting (FACS) to obtain cell populations enriched in adult and fetal mammary stem cells. Gene expression data for these FACS enriched cell populations were obtained from the following published studies involving three human and two murine analyses: GSE16997 [1], GSE19446 [2], GSE27027 [3], GSE30489 [4], GSE35399 [5]. Using similar methods and cell markers, an additional human dataset was also newly derived by the Perou Lab[6]. To identify genes uniquely and highly expressed within each enriched mammary cell population, including the fMaSC fraction, a two-class Significance Analysis of Microarrays (SAM) [7, 8] was performed within each dataset comparing each given FACS population versus all others from that experimental data set to generate a total of 23 different population signatures. We have now tested all 23 signatures for prognostic significance, and for their ability to predict neoadjuvant chemotherapy response.

We found that, typically, genes within an individual signature were highly correlated within their respective cell population. For example, a FACS fraction defined functionally by *in vitro* colony formation as representing mature luminal cells exhibited significant differential expression of genes expected to be associated with this biological context (i.e. K8, ESR1, GATA3, etc). This is an important confirmation of the classifier strategy employed. However, we occasionally observed expression of some genes that one might expect to derive from cell types not specifically designated by a given FACS profile. This does not necessarily indicate similar mechanisms of gene regulation, nor similar biological consequences. For example, expression of luminal and myoepithelial specifier genes within a FACS population may either result from representation of luminal and myoepithelial cells within that cell population, or, alternatively, expression of both types of genes within individual cells that are bi-potent progenitors or stem cells. While the single cell sequencing studies being performed in Aim 2 can resolve these possibilities, in Aim 1 we attempted to use a bioinformatics approach to identify “refined sub-signatures” within each of the FACS derived cell populations to obtain reduced gene sets that might be more clinically robust than their broader parental signatures. We used the UNC308 breast tumor dataset [9] to hierarchically cluster genes derived from each FACS population separately, and then identified sub-signatures referred to as “refined1, refined2, etc.”, as a gene sub-set having at least ten genes AND exhibiting an intra-cluster Pearson correlation greater than 0.5 (see Hoadley et al 2007[10], for details of this method). This enabled us to refine the fMaSC signature into three sub-signatures. As we described previously, the entire fMaSC signature is most significantly enriched in basal-like tumors [3] (Figure 1A). By contrast, the three sub-signatures varied significantly across intrinsic breast cancer subtypes. Thus, fMaSC-refined1 is highest in basal-like tumors, fMaSC-refined2 is similarly expressed across the subtypes, and fMaSC-refined3 is expressed in luminal tumors (Figure 1B). These results are consistent with our hypothesis that subsets of genes within the parental fMaSC signature are likely regulated by different biological mechanisms, increasing the importance of finding sub-signatures for developing robust clinical tests.

To determine if any of the FACS enriched signatures, especially the fMaSC-refined signatures, are prognostic and whether they can predict response to chemotherapy, we tested them using a 441 patient dataset from the MD Anderson Cancer Center [11] (GSE25066), which has both relapse free survival (RFS) and pathologic complete response (pCR) data to

anthracycline/taxane containing neoadjuvant chemotherapy. A univariate analysis for RFS identified several prognostic signatures, including the fMaSC-refined1 (Odds Ratio (OR)=1.2, p-value=0.036) and fMaSC-refined3 (OR =0.6, p-value<0.001). When these signatures are used in a univariate pCR analysis, fMaSC-refined1 (OR =2.54, p-value<0.001) and fMaSC-refined3 (OR =0.46, p-value<0.001) were again significant for determining patient subgroups that are likely to respond to chemotherapy across all breast cancer subtypes (Figure 1C). While these results are interesting, it is known that basal-like tumors are more likely to respond to chemotherapy than the luminal subtypes [12]. As such, we also performed a multivariate analysis in which we included subtype, ER receptor status, PR receptor status, tumor stage, tumor grade, and proliferation score in the model to determine if our refined-signatures improve upon current clinical markers. In this analysis, the fMaSC-refined1 sub-signature remained highly significant (OR =2.07, p-value<0.001). Therefore, we suggest that the fMaSC refined1 sub-signature contains genes or biological processes independent of known factors that correlate with chemotherapy response. This suggests that biological processes unique to fMaSCs, a biologically defined cell population not used previously to deduce prognostic signatures, are associated with the chemotherapy sensitivity of basal-like tumors. This justifies our proposal to further characterize fMaSCs using either purification or single cell sequencing strategies, followed by derivation of more precise gene signatures from this important cell type. In addition, we will continue to explore orthogonal approaches to define cancer relevant sub-signatures, deduce the biological processes they engender and identify therapeutic targets.

Aim 2

We have significantly improved our ability to obtain RNA sequence from single cells from the fMaSC-enriched population over the past year. We obtained new single cell RNA sequencing data on 15 fMaSC candidate cells defined by their co-expression of Krt8, Krt18, Krt14, and EpCAM among other genes. These data provide a precedent for the likely success of our proposal to use single cell sequencing to elucidate gene control modules that underlie the stem cell state of these early embryonic cells, and for elucidating genes and pathways they share with basal-like human breast cancers (see Figure 2). However, we also encountered, and have now overcome, significant challenges to achieving this Aim. The information we gained in the first year, along with new technical approaches and instrumentation, position us well to complete our objectives in the second year of this Idea Expansion Award.

Our initial RNA sequencing produced an unexpected result. Unlike our previous immunofluorescence-based analysis that showed that only 30-40% of the cells at embryonic day 18.5 (E18.5) co-express K14 and K8, RT-PCR analysis of our RNA-sequencing libraries showed that most E18.5 epithelial cells are K14/K8 double positive at the RNA level. This presents a problem regarding our initial proposal to sequence K14 and K8 mono-positive cells from E18.5 to use as comparators for cells that had undergone, or were undergoing, myoepithelial and luminal differentiation, respectively. Since overtly lineage committed cells turn out to be rare at E18.5, we now intend to pursue the following novel alternative to identify differentially expressed, stem-cell specific genes. We will compare the transcriptional profiles of single cells from E18.5 fetal mammary rudiments, which are enriched in fMaSC activity, to cells from E15.5, which we previously showed lack detectable fMaSC activity as measured by the gold standard of *in vivo* transplantation. This approach will not only enable us to identify biomarkers useful for prospectively identifying fMaSCs, but should also elucidate the changes in gene expression that occur upon the generation of fMaSCs from their developmental antecedents.

We will also determine the gene expression characteristics of the diverse stromal cells that associate with the fMaSCs before and after they become specified. This is critical as stem cells require interaction with a specific microenvironment, or 'niche'. We expect that such interactions involve interaction of the fMaSCs with specific paracrine factors secreted by the stroma and via fMaSC-encoded receptors. We already have evidence for this in terms of the requirements for EGF, FGF, and HGF for optimal fMaSC activity. Further, we suspect that there may be multiple cell types that constitute the fMaSC niche, and that cell-cell and cell-matrix associations may also be important, as indicated by the expression of specific integrins on the fMaSCs. We will, therefore, obtain a higher resolution analysis of the stromal component by isolating the stroma (non-epithelial cells), and sequencing individual cells to identify potential paracrine activators of the mammary stem cell state. We expect this to have relevance for understanding growth regulatory mechanisms of triple-negative breast cancers, as fMaSCs and TNBCs of the Basal-like subtype have highly related gene expression signatures. Moreover, our previous microarray data justify the broader importance of this approach since we found significant prognostic value associated with stromal gene expression signatures, such as those associated with a wound response evident in fMaSC-associated stroma.

Previously, technical and cost-considerations made such proposals untenable. However, much has changed in the past year to make these objectives attainable. First, we have significantly improved the reproducibility, reliability, and speed with which we can obtain single cell sequencing data. Although we were able to obtain sequencing data from single fMaSC candidate cells, the high variance in expression scores compromised our ability to confidently identify differentially expressed genes, clustering of cells into cell types, and deduction of stem cell signatures. While significant variance in expression levels is a recognized feature of single cell data, the relatively low mapping frequencies we obtained further undercut confidence in overall quality of the data. Over the past year, the Lasken lab has been refining single cell sequencing methods to improve data quality and consistency. This led us to switch to the Smarter technology (Clontech) for cDNA synthesis, which affords superior performance compared to the older Life Technologies' method. We also switched to the Illumina HiSeq platform from the Life Technologies SOLiD platform that we originally proposed. The reason for this is that the new Illumina method provides superior read quality and simpler analysis of data. We have also developed and implemented new critical quality control measures at each point in library preparation including incorporation of "spike in controls" that provide a parallel efficiency measure at each point throughout the process from cDNA synthesis through sequencing and mapping.

We have also identified major sources that contribute noise to the data, and have taken steps to alleviate such issues. First, we found that manual cell manipulation including pipetting can generate significant variations within and between experiments. Manual micro-manipulation of large numbers of cells, such as we originally carried out, is laborious and exposes the cells to potential damage and stresses that could perturb gene expression patterns during the long times required for single cell isolation. Manual pipetting of limiting sample quantities is highly susceptible to person-to-person and sample-to-sample variability as well. To mitigate these concerns, we have adapted the automated C1 microfluidic device (Fluidigm) to rapidly obtain cDNA libraries from individual cells. This device enables up to 96 cells to be captured in isolated microfluidic wells in a fraction of the time required for manual micromanipulation, and the most sensitive early stages of cDNA library preparation are then carried out in parallel directly on the chip, which eliminates losses associated with manual sample preparation methods. Using this pipeline with the new library preparation and quality control approaches described above, we now achieve mapping efficiencies >95%, as opposed compared to <50% from our prior method.

The C1 approach also enables us to rapidly isolate individual cells that we prescreen for viability to ensure that each RNA-seq library derives from a viable cell (Figure 2). We confirmed in a recent study employing the C1 that the vast majority of cells from E18.5 do indeed co-express K14 and K8 (Figure 2). These data provide strong rationale for analyzing both fMaSCs and their associated stroma, and for comparing gene expression data from E15.5 (little or no fMaSC activity) and E18.5 (high fMaSC activity). Such studies are now feasible, and within the budget proposed, due to our methodological improvements and the significant decreases in sequencing costs over the past year. Importantly over the past year, the Salk Institute has added a sequencing core and significant bioinformatics personnel and resources. Therefore, we will now do the sequencing at the Salk Institute instead of the J. Craig Venter Institute, as this will be more convenient technically and more cost effective.

Figure 1: Analysis of the fMaSC signature, and refined sub-signatures. Box and whisker plots showing the average expression of the complete fMaSC signature (A), and three refined sub-signatures (B) are plotted according to breast tumor intrinsic subtype using 308 primary breast tumors. (C) Multivariate analysis predicting pathological complete response to neoadjuvant chemotherapy was performed using 441 patients, with the significant features shown in grey shading. Note that fMaSC-refined1 is significant even after accounting for ER status, stage, grade, and PAM50 intrinsic subtype.

Figure 2: Transcript reads from preliminary sequencing of 15 candidate fMASC cells aligned to gene models for Krt14, Krt8, Krt18 and EpCam illustrating the feasibility of sequencing individual fMaSC cells and verifying co-expression of basal and luminal keratins.

Figure 3: A revised experimental pipeline for RNA-Seq on single fMaSC and related cells. (A) high power confocal image of two of the 96 Fluidigm-C1 capture wells containing candidate fMaSC cells. green=live (Calcein-AM), red=dead (Ethidium Bromide). (B) RT-PCR analysis of prepared cDNA libraries indicating the capture and processing of >50 live keratin double positive cells. (C) Expected distribution of Nextera (Illumina) library fragment size from individual samples indicating production of high quality, single cell RNA Sequencing libraries.

Key Research Accomplishments:

Aim 1

- Obtained/derived transcriptomic profiles of six mammary cell populations across four human and two mouse datasets, with a focus on the fMaSC signature
- Discovered sub-signatures within the parental fMaSC signature using a diverse human breast cancer dataset.
- Performed univariate and multivariate RFS and pCR chemotherapy response testing on all derived gene signatures using the 441 patient MD Anderson dataset.
- Identified the fMaSC-refined1 sub-signature as highly predictive of chemotherapy response even when controlling for common clinical variables across all patients.
- Profiled and credentialed 27 distinct mouse mammary tumor models for model selection for future tumor and fMaSC studies; this genomic study is now In Press[13], and acknowledges this grant.

Aim 2

- Investigated and optimized methods for obtaining single viable embryonic cells enriched in fMaSCs
- Investigated and optimized methods to make libraries for RNA-sequencing from single fMaSCs
- Obtained RNA-seq data from 15 K14/K8 double positive cells using the Illumina platform and have gained expertise analyzing RNA-seq data
- Showed by single cell RNA-seq that the majority of cells within the fMaSC-enriched population at E18.5 express K14 and K8
- Used the Fluidigm C1 microfluidics platform to isolate >50 viable K14/K8 cells from E18.5, prepared cDNA and genomic libraries for RNA-seq.

Conclusions:

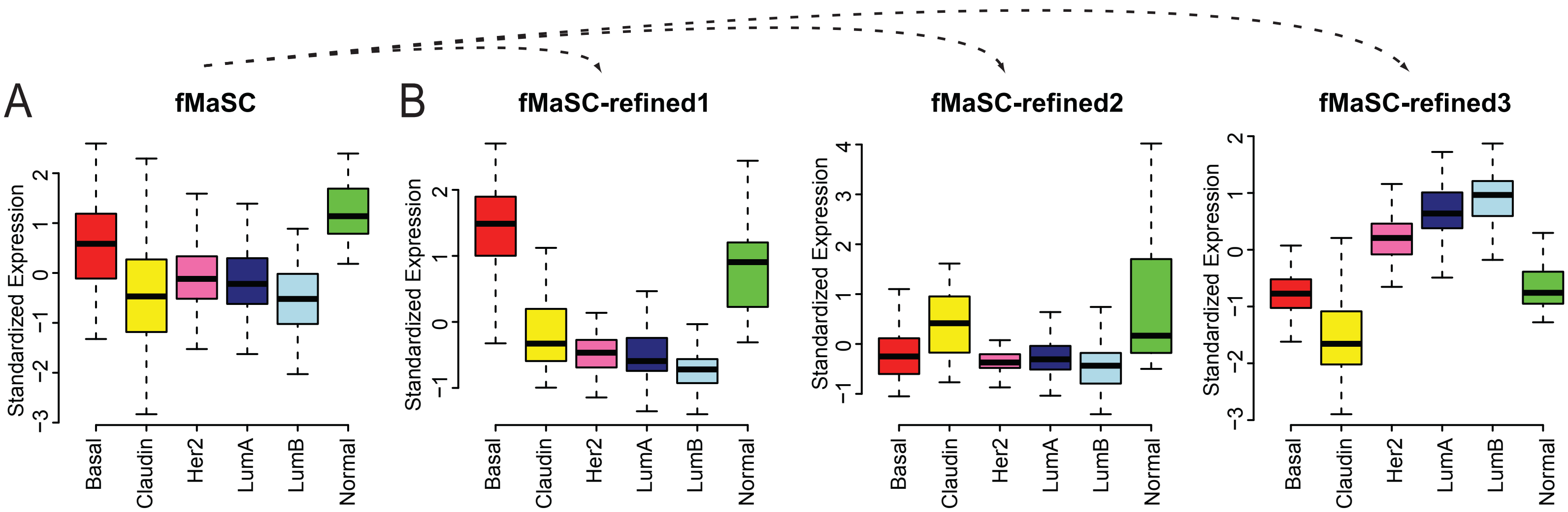
The finding that the fMaSC-refined1 sub-signature is highly predictive of chemotherapy response under rigorous multivariate testing conditions is an exciting result. This finding suggests that this signature could be important for helping clinicians determine which subsets of breast cancer patients are most likely to respond to standard anthracycline and taxane chemotherapy treatment regimens. We are currently following up on these results by seeing if they can be verified using the recently published Horak *et al* [14] dataset.

In addition, we also propose to perform three sets of experiments to ascertain the generality of our preliminary findings. First, we will continue with our bioinformatics approaches to evaluate other methods of gene list refinement, and to “refine” other FACS population signatures, mostly importantly the adult Mammary Stem Cell signature. Second, we will perform immunofluorescence and/or RNA *in situ* hybridization on adult and fetal mammary tissues to determine what morphologically distinct cell types are making the genes contained with fMaSC-refined1 and fMaSC-refined3. Third, we will perform single cell RNA-seq on selected cells from the fMaSC- and related populations to determine if single cells simultaneously express both the fMaSC-refined1 and fMaSC-refined3 sub-signatures, or if these signatures define two different cell types and to define markers and regulatory mechanisms contributing to the fMaSC phenotype.

References

1. Lim E, Vaillant F, Wu D, Forrest NC, Pal B, Hart AH, Asselin-Labat ML, Gyorki DE, Ward T, Partanen A, et al: **Aberrant luminal progenitors as the candidate target population for basal tumor development in BRCA1 mutation carriers.** *Nat Med* 2009, **15:**907-913.
2. Lim E, Wu D, Pal B, Bouras T, Asselin-Labat ML, Vaillant F, Yagita H, Lindeman GJ, Smyth GK, Visvader JE: **Transcriptome analyses of mouse and human mammary cell subpopulations reveal multiple conserved genes and pathways.** *Breast Cancer Res* 2010, **12:**R21.
3. Spike BT, Engle DD, Lin JC, Cheung SK, La J, Wahl GM: **A mammary stem cell population identified and characterized in late embryogenesis reveals similarities to human breast cancer.** *Cell Stem Cell* 2012, **10:**183-197.
4. Keller PJ, Arendt LM, Skibinski A, Logvinenko T, Klebba I, Dong S, Smith AE, Prat A, Perou CM, Gilmore H, et al: **Defining the cellular precursors to human breast cancer.** *Proc Natl Acad Sci U S A* 2012, **109:**2772-2777.
5. Shehata M, Teschendorff A, Sharp G, Novcic N, Russell A, Avril S, Prater M, Eirew P, Caldas C, Watson CJ, Stingl J: **Phenotypic and functional characterization of the luminal cell hierarchy of the mammary gland.** *Breast Cancer Res* 2012, **14:**R134.
6. Prat A, Karginova O, Parker JS, Fan C, He X, Bixby L, Harrell JC, Roman Perez E, Troester M, Perou CM: **Characterization of cell lines derived from breast cancers and normal mammary tissues for the study of the intrinsic molecular subtypes.** *Breast Cancer Research and Treatment* In Press.
7. Prat A, Parker JS, Karginova O, Fan C, Livasy C, Herschkowitz JI, He X, Perou CM: **Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer.** *Breast Cancer Res* 2010, **12:**R68.
8. Tusher VG, Tibshirani R, Chu G: **Significance analysis of microarrays applied to the ionizing radiation response.** *Proc Natl Acad Sci U S A* 2001, **98:**5116-5121.
9. Harrell JC, Prat A, Parker JS, Fan C, He X, Carey L, Anders C, Ewend M, Perou CM: **Genomic analysis identifies unique signatures predictive of brain, lung, and liver relapse.** *Breast Cancer Res Treat* 2012, **132:**523-535.
10. Hoadley KA, Weigman VJ, Fan C, Sawyer LR, He X, Troester MA, Sartor CI, Rieger-House T, Bernard PS, Carey LA, Perou CM: **EGFR associated expression profiles vary with breast tumor subtype.** *BMC Genomics* 2007, **8:**258.
11. Hatzis C, Pusztai L, Valero V, Booser DJ, Esserman L, Lluch A, Vidaurre T, Holmes F, Souchon E, Wang H, et al: **A genomic predictor of response and survival following taxane-anthracycline chemotherapy for invasive breast cancer.** *JAMA* 2011, **305:**1873-1881.
12. Carey LA, Dees EC, Sawyer L, Gatti L, Moore DT, Collichio F, Ollila DW, Sartor CI, Graham ML, Perou CM: **The triple negative paradox: primary tumor chemosensitivity of breast cancer subtypes.** *Clin Cancer Res* 2007, **13:**2329-2334.
13. Pfefferle AD, Usary J, Herschkowitz J, Darr D, Harrell JC, Spike B, Wahl G, Rosen JM, Perou CM: **Transcriptomic classifiation of genetically engineered mouse models of breast cancer identifies human counterparts.** *Genome Biology* In Press.
14. Horak CE, Pusztai L, Xing G, Trifan OC, Saura C, Tseng LM, Chan S, Welcher R, Liu D: **Biomarker analysis of neoadjuvant doxorubicin/cyclophosphamide followed by ixabepilone or Paclitaxel in early-stage breast cancer.** *Clin Cancer Res* 2013, **19:**1587-1595.

Figure 1



C

	Patients	Univariate p-value	Univariate Odds Ratio	Multivariate p-value	Multivariate Odds Ratio
fMaSC	441	0.530	1.08 (0.85-1.39)	0.239	1.22 (0.88-1.72)
fMaSC-refined1	441	<0.001	2.54 (1.98-3.32)	<0.001	2.07 (1.41-3.13)
fMaSC-refined2	441	0.909	1.01 (0.80-1.29)	0.891	1.02 (0.77-1.36)
fMaSC-refined3	441	<0.001	0.46 (0.35-0.60)	0.688	1.10 (0.68-1.79)
PAM50 Proliferation	441	<0.001	2.52 (1.89-3.47)	0.005	1.91 (1.23-3.03)
Estrogen Receptor (ER)					
negative	175 (40%)		1		1
positive	266 (60%)	<0.001	0.23 (0.14-0.39)	0.711	0.86 (0.39-1.87)
Progesterone Receptor (PR)					
negative	227 (51%)		1		1
positive	214 (49%)	<0.001	0.30 (0.18-0.51)	0.980	0.99 (0.46-2.14)
Tumor Stage					
1	27 (6%)		1		1
2	226 (51%)	0.364	0.65 (0.27-1.75)	0.447	0.66 (0.23-2.03)
3	126 (29%)	0.747	0.85 (0.34-2.36)	0.387	0.61 (0.20-1.93)
4	62 (14%)	0.054	0.31 (0.09-1.02)	0.019	0.20 (0.05-0.76)
Tumor Grade					
1	28 (6%)		1		1
2	170 (39%)	0.500	2.05 (0.38-38.1)	0.782	1.36 (0.22-26.6)
3	243 (55%)	0.019	11.1 (2.3-200.7)	0.334	2.92 (0.48-56.9)
PAM50 Subtype					
Luminal A	141 (32%)		1		1
Luminal B	67 (15%)	0.003	6.01 (1.92-22.63)	0.356	1.92 (0.50-8.41)
HER2-enriched	24 (5%)	0.010	6.85 (1.51-31.13)	0.207	2.77 (0.55-14.1)
Basal-like	167 (38%)	<0.001	18.2 (7.21-61.46)	0.038	4.28 (1.15-18.8)
Normal-like	42 (10%)	0.001	8.06 (2.39-31.68)	0.005	7.04 (1.89-29.8)

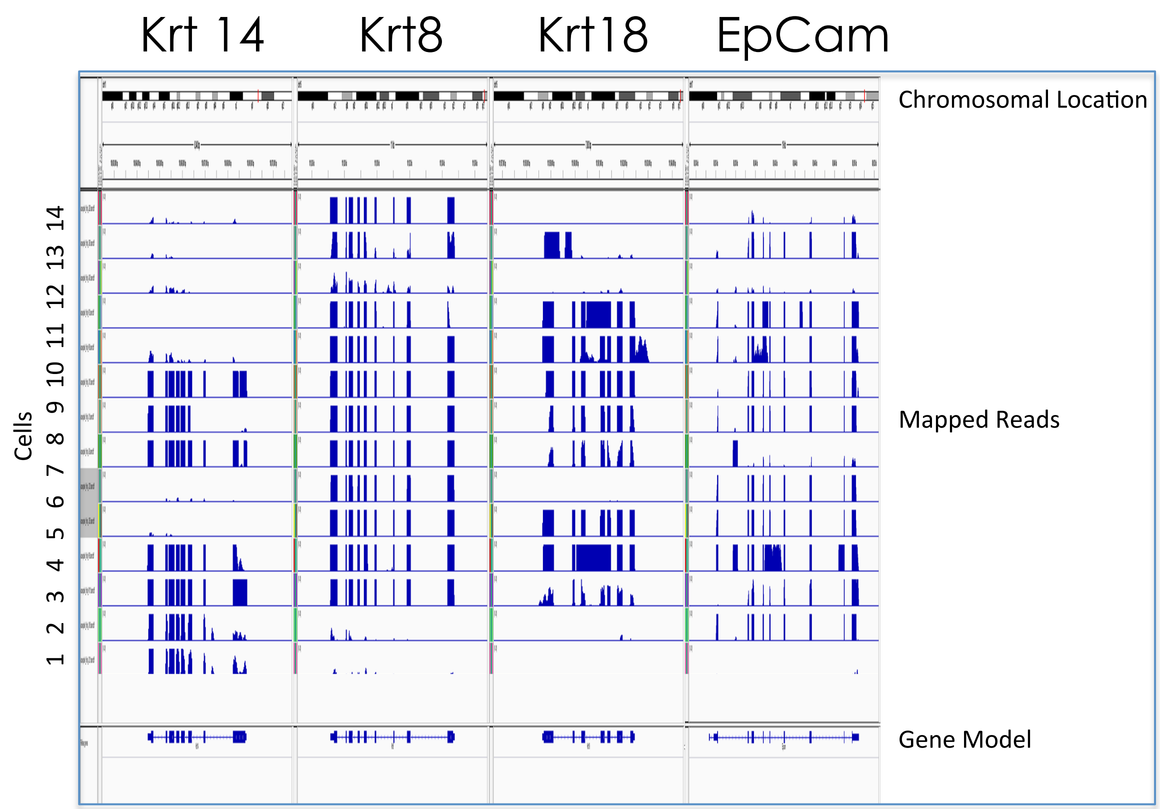


Figure 2: Reads from preliminary sequencing of 15 candidate fMASC cells aligned to gene models for Krt14, Krt8, Krt18 and EpCam illustrating the feasibility of sequencing individual fMaSC cells and verifying co-expression of basal and luminal keratins.

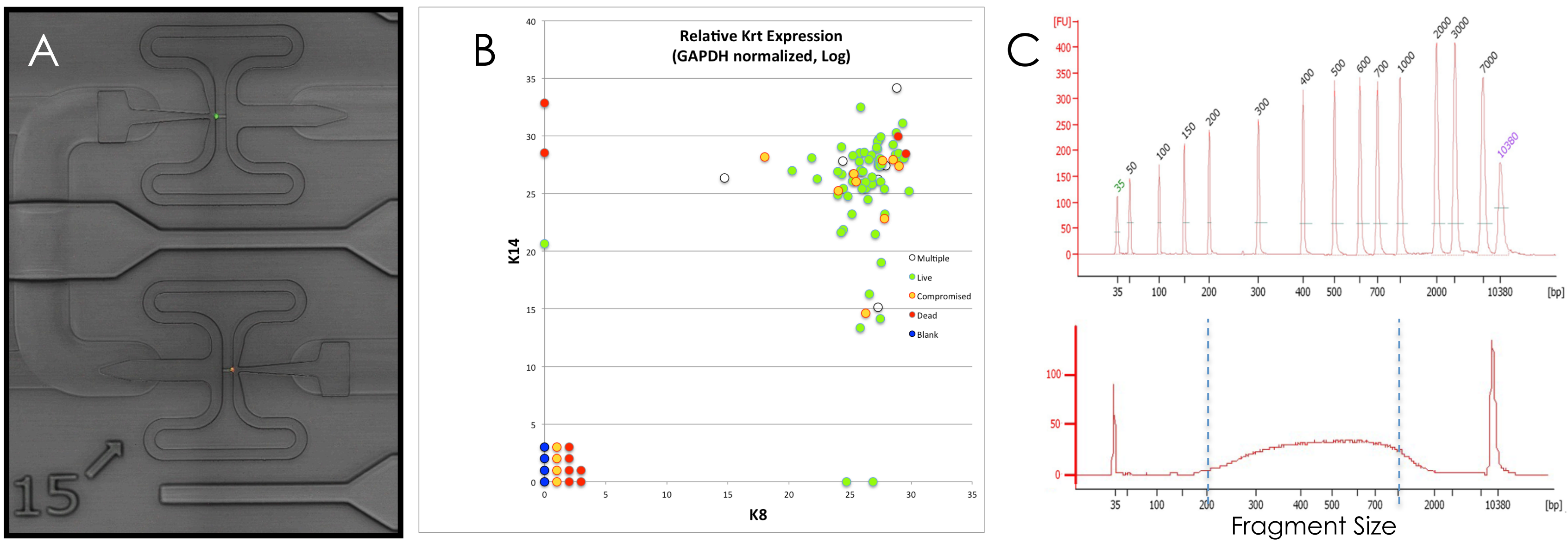


Figure 3: A revised experimental pipeline for RNA-Seq on single fMaSC and related cells. (A) high power confocal image of two of the 96 Fluidigm-C1 capture wells containing candidate fMaSC cells. green=live (Calcein-AM), red=dead (Ethidium Bromide). (B) RT-PCR analysis of prepared cDNA libraries indicating the capture and processing of >50 live keratin double positive cells. (C) Expected distribution of Nextera (Illumina) library fragment size from individual samples indicating production of high quality, single cell RNA Sequencing libraries.